

REGEN: A Dataset and Benchmarks with Natural Language Critiques and Narratives

Kun Su*
Google Research
Mountain View, CA, USA

James Pine*
Google Research
Mountain View, CA, USA

Marialena Kyriakidi
Google Research
Mountain View, CA, USA

Anushya Subbiah
Google Research
Mountain View, CA, USA

Krishna Sayana*
Google Research
Mountain View, CA, USA

Yuri Vasilevski*
Google Research
Mountain View, CA, USA

Liam Hebert†
University of Waterloo
Waterloo, Ontario, Canada

Hubert Pham*
Google Research
Mountain View, CA, USA

Raghavendra Vasudeva
Google Research
Mountain View, CA, USA

Ambarish Jash
Google Research
Mountain View, CA, USA

Sukhdeep Sodhi
Google Research
Mountain View, CA, USA

Abstract

This paper introduces a novel dataset REGEn (Reviews Enhanced with GENERative Narratives), designed to benchmark the conversational capabilities of recommender Large Language Models (LLMs), addressing the limitations of existing datasets that primarily focus on sequential item prediction. REGEn extends the Amazon Product Reviews dataset by inpainting two key natural language features: (1) user critiques, representing user "steering" queries that lead to the selection of a subsequent item, and (2) narratives, rich textual outputs associated with each recommended item taking into account prior context. The narratives include product endorsements, purchase explanations, and summaries of user preferences.

Further, we establish an end-to-end modeling benchmark for the task of conversational recommendation, where models are trained to generate both recommendations and corresponding narratives conditioned on user history (items and critiques). For this joint task, we introduce a modeling framework LUMEN (LLM-based Unified Multi-task Model with Critiques, Recommendations, and Narratives) which uses an LLM as a backbone for critiquing, retrieval and generation. We also evaluate the dataset's quality using standard auto-rating techniques and benchmark it by training both traditional and LLM-based recommender models. Our results demonstrate that incorporating critiques enhances recommendation quality by enabling the recommender to learn language

understanding and integrate it with recommendation signals. Furthermore, LLMs trained on our dataset effectively generate both recommendations and contextual narratives, achieving performance comparable to state-of-the-art recommenders and language models.

CCS Concepts

• Information systems → Recommender systems; • Computing methodologies → Natural language generation.

Keywords

Sequential Recommenders, Large Language Models, Generative Recommenders

ACM Reference Format:

Kun Su, Krishna Sayana, Hubert Pham, James Pine, Yuri Vasilevski, Raghavendra Vasudeva, Marialena Kyriakidi, Liam Hebert, Ambarish Jash, Anushya Subbiah, and Sukhdeep Sodhi. 2018. REGEn: A Dataset and Benchmarks with Natural Language Critiques and Narratives. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 11 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Applying Large Language Models (LLMs) to conversational recommenders [9, 27, 35, 50, 53] has received some interest of late. In addition to providing personalized recommendations, conversational recommenders engage in natural dialog with users e.g. to explain recommendations, refine recommendations based on natural language feedback from the user. LLMs show promise in fulfilling the necessary language capabilities of such recommenders.

This is challenging as existing datasets for training conversational recommender systems lack rich, contextual narratives and explicit user feedback. Such a dataset is crucial for modeling natural language interactions in the context of the recommended items, so generated text is consistent and relevant to the recommendations and user preferences. While several recent works explore modeling approaches that integrate the recommender's internal state with

*Equal contribution.

†Work done while at Google Research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06

<https://doi.org/XXXXXXX.XXXXXXX>

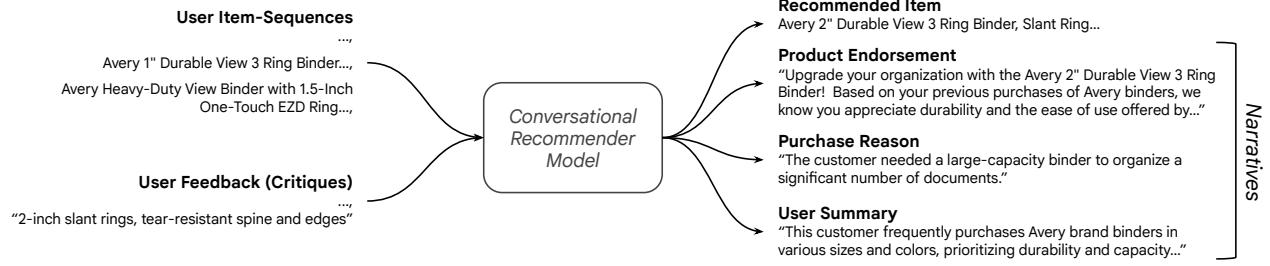


Figure 1: REGEN is a dataset for training conversational recommender models that output both item recommendations and engaging natural language, given user history and feedback. Based on the user item-sequences in Amazon Product Reviews, REGEN adds (1) inferred user feedback, in the form of natural language critiques, and (2) rich text narratives, in the form of product endorsements, purchase reasons, and user summaries—all personalized to each user.

an LLM [16, 36, 50, 53], they are limited to existing datasets (e.g. [5, 9, 25, 27, 34, 47]) that concentrate on user item-sequences, short dialog snippets, structured outputs, or small summaries that lack the rich and varied conversational elements.

Collecting such a dataset from users is challenging, and would typically require a large set of real users interacting with a conversational recommender. We thus continue the trend [6, 23, 34] of exploring the hypothesis that LLMs can (1) infer user preferences and goals from an existing dataset of user behavior traces; and (2) generate compelling and plausible user narratives and responses, consistent with ground truth behavior. Aside from expediency, an advantage to this approach is that we can augment existing popular sequential recommendation datasets.

We present REGEN (Reviews Enhanced with GEnerative Narratives), a dataset that augments the Amazon Reviews datasets [34] by incorporating natural language outputs relevant to conversational recommendations (Figure 1). REGEN adds diverse narratives such as purchase reasons, explanations, product endorsements, user summaries & concise user profiles, personalized to each user. It also adds plausible critiques, or natural language feedback that the user issues in response to recommendations to steer the recommender. To augment the dataset, we use Gemini 1.5 Flash [44] with few shot prompting to inpaint narratives and critiques. The LLM generates those features using the user’s rating history, product review text, and item descriptions in the original Amazon Reviews datasets. As an initial measure of feasibility, we automate quality inspection with an auto-rater LLM [10]. The auto-rater assesses the generated outputs across multiple attributes, with a focus on grounding and factuality based on the user’s historical interactions.

In addition to a dataset, there is also a need for a single training task that integrates user behavior and recommendations, to evaluate a model’s consistency and handling of those aspects. While many previous studies (e.g. [9, 36]) propose models that are trained with separate individual objectives, e.g. to make recommendations, rating predictions, or review generation, we believe a task that combines those aspects is a critical step towards evaluating coherence in conversational recommendations. To aid model development, we propose a single-turn conversational task as a step towards fully conversational recommenders. In this task, the model input is a user history of items, combined with user critiques between some items. The output is the next item to recommend and rich

natural language narratives consistent with the recommend item and user preferences. The task requires the model to effectively combine personalized sequential recommendation, critiquing, and natural language generation into a single, coherent turn of a fully conversational recommender.

Finally, to evaluate the utility of both the REGEN dataset and task, we present reference models and strong baselines. Here we introduce LUMEN (LLM-based Unified Multi-task Model with Critiques, Recommendations, and Narratives), a single model using an LLM backbone trained to generate recommendations and narratives by integrating collaborative filtering signals, content signals, and natural-language critiques inputs. LUMEN leverages user-item interaction data, rich item content representations, and user feedback to generate outputs that are informative and aligned with user preferences. Our experiments establish benchmarks that demonstrate the effectiveness of this architecture in producing human-like conversational recommendations.

In summary, the key contributions are:

- REGEN, an open-source dataset,¹ with critiques and rich natural language narratives that are consistent and personalized to a user’s consumption history. We evaluate it for factuality, grounding, and accurately reflecting user context.
- A corresponding task towards a fully conversational recommender which can serve as a foundation for future research.
- Reference models that show LLM architectures combining generative recommendation and language task achieve performance comparable to state-of-the-art recommenders and language models.
- Validation of the model’s ability to learn and generate contextually rich narratives and leverage user critiques to dynamically guide recommendations. We further demonstrate generation of coherent conversational traces.

2 Related Work

REGEN is inspired by existing datasets and modeling tasks that support the development and evaluation of recommender systems.

¹<https://www.kaggle.com/datasets/googleai/regen-reviews-enhanced-with-generative-narratives>

2.1 Datasets

2.1.1 Conversational. Many conversational datasets are dialog-based, representing conversation between a user and an agent. For example, DuRecDial [29, 30] is a crowd-sourced, multi-domain dialogue dataset with pre-defined goals. E-ConvRec [19] provides a large-scale conversational recommendation dataset based on real pre-sales dialogues between users and customer service. ReDial [24] is an annotated dataset of dialogues of movie recommendations. TG-ReDial [52] extends ReDial by incorporating topic information to guide conversations and recommendations, while E-Redial [12] expands it with external knowledge to improve explainability. These datasets focus on constructing or collecting dialogues, but typically lack coupling between the dialogue and the user’s underlying preferences or historical actions. LLM-Redial [27] pursues a user-centric focus by creating dialogues via templates, rooted in real user profiles. REGEN is similar to LLM-Redial with its focus on maintaining consistency between the user profile and the narratives but differs in several ways. First, REGEN preserves the sequential item history to maintain temporal consistency, enabling it to generate user feedback and narratives conditioned on past user behavior. Second, rather than short dialogue turns, REGEN focuses on personalized, rich narratives that a knowledgeable conversational recommender might produce. We believe a narrative-focused dataset helps test a recommender’s language understanding, in the context of the recommended items and user preferences.

2.1.2 Narratives. The “Justifying Recommendations” dataset by Ni et al. [34] and the “Unlocking the Why of Buying” by Chen et al. [5] focus on generating explanations from reviews. The “Amazon-C4” dataset by Hou et al. [15] presents a new reviews dataset that constructs complex queries and associates them with items to evaluate the models on recommendation tasks. REGEN extends their work with open-ended narratives suited to several recommendation tasks, conditioned on the entire user history, and covering use cases beyond purchase explanations. REGEN targets natural language outputs that reflect the rapidly evolving capabilities of LLMs, with longer, more nuanced, and semi-structured and unstructured responses.

2.1.3 Critiquing. Critiquing for classical recommender systems is surveyed in Chen and Pu [4], Gao et al. [8], Jannach et al. [18]. Early critiquing systems enabled the user to provide feedback on recommended items, typically at the feature-level, through (e.g.) suggested pre-designed tweaks [2] or language-driven dialog systems [11, 41, 46]. Recent work [31, 32, 48] explores systems that jointly learn latent representations for users, items, and key-phrases, the latter of which serves as critiques that can influence the recommendations through operations in the latent space. Others explore the interpretation of attributes [1, 33] to support critiquing. These works provide the basis for natural-language critiques as a vehicle for the user-feedback that REGEN provides.

2.2 Recommendation Tasks

2.2.1 Sequential Item Recommendations. There is a plethora of ongoing research in an effort to combine LLMs with recommender systems for sequential recommendations [3, 7, 13, 14, 21, 26, 28, 36–38, 40, 42, 49, 51, 54]. These models explore different approaches

for representing item and user interactions (e.g. ID embeddings, text embedddings), while some of them seek to enhance quality by incorporating item and/or semantic context. Item recommendation is an important, primary task that REGEN supports. As it is based on the Amazon Reviews dataset, a popular benchmarking dataset, we believe REGEN is a natural fit for extending existing LLM-based sequential recommenders to also generate language.

2.2.2 Item Recommendation and Text Generation. There are a number of recommendation models that output both item recommendations and natural language. P5 [9] represents all user-item signals, along with associated context, as natural language sequences that are input to an encoder/decoder model. It proposes several task families for item prediction and text generation. User-LLM [36] encodes user history into user embeddings, and incorporates them into an LLM via cross-attention. User-LLM supports three tasks, next item recommendation, genre prediction and reviews generation. While those projects propose tasks that involve both recommendations and text generation, they are separate tasks. In contrast, REGEN proposes integrating both those tasks, along with incorporating user feedback, into a single coherent task, while generating a rich and diverse set of narratives.

3 REGEN

This section details the REGEN dataset, including the generation process for the critiques and narratives, and evaluation methodology with a rater LLM. Further details of the modeling tasks for benchmarking are described in the following section.

3.1 Amazon Reviews

We use datasets from Amazon Product Reviews [34] as a basis for augmentation with rich narratives. The Amazon Reviews dataset (2018) is a massive collection of customer reviews spanning various product categories. This version, released in 2018, contains over 233 million reviews, making it one of the largest publicly available datasets for sentiment analysis and recommendation systems. Each review includes item features like title, description, category, price and review features including text, timestamp, score and summary. User sequences can be created by sorting each user review by timestamp, creating a sequential recommendation task. For our work, we use “Office Products” and “Clothing, Shoes and Jewelry,” verticals. These are chosen to be representative samples with different item counts, Office with 27k items and Clothing with 376k items.

Table 1: Statistics for the CALRec-processed Amazon Product Reviews dataset. I/U is the average number of items per user, or average sequence length.

Dataset	# Users	# Items	I/U
Office	89,489	27,887	7.86
Clothing	1,168,735	376,847	9.21

3.2 Dataset Generation

3.2.1 Data Preprocessing: User sequences $S_u = \{i_1, i_2, \dots, i_n\}$ consisting of interactions i are first created by aggregating reviews

per user, and sorting them by timestamp. For comparable results, we use the CALRec [26] preprocessed datasets. Notably, CALRec's de-duplication methodology helps mitigate trivial recommendation strategies, e.g. always recommending the most recent item. See Table 1 for the dataset statistics.

3.2.2 Inpainting for Synthesizing Multi-Turn Conversations: The Amazon Reviews dataset, while valuable for recommendation studies, primarily reflects a series of individual user interactions rather than real conversations. To create a synthetic dataset that emulates multi-turn conversational dynamics, we employ an inpainting technique to generate the elements typically missing in non-conversational review data. Specifically, we leverage LLMs to synthesize these elements, using carefully designed prompts and an iterative evaluation process to ensure high-quality generation and minimize inaccuracies. This allows us to compare against existing research on these datasets for the recommendation task while adding natural language elements and setting new baselines for the language task. In this work, we utilize the Gemini 1.5 Flash model, a fast and efficient version of the Gemini 1.5 model family [44].

We exploit the extended context length capabilities of the LLM to effectively process the complete user history S_u for each user u . Our prompts are simple, employing a task prefix with instructions and the desired output format. This prefix is followed by the user's entire interaction history, including item metadata and review text. Finally, the prompt specifies the expected output format. We use the user history and associated reviews to generate critiques, purchase reasons and summaries. For product endorsement generation, however, we intentionally withhold the user's last review, enabling the model to learn how to endorse a recommended item without being influenced by that item's review.

3.2.3 Critiques Generation. REGEN adds generated critiques (short utterance) to steer the system from the current recommended item to a desired item. Our aim is to focus on the setting where a user refines from one item to a closely-related item variant (e.g., “red ball-point pen” to “black ball-point pen”), rather than to another arbitrary item, which would be better served by other mechanisms, e.g. a new search query. Thus, REGEN only generates critiques between adjacent pairs of items that are sufficiently similar. We use the Amazon Reviews-provided hierarchical item categories as a rough proxy for item similarity and consider items sufficiently similar if at least four levels of the category match.

To generate critiques for adjacent item pairs that meet the similarity criteria, we query Gemini 1.5 Flash to synthesize several options, instructing the LLM to treat the first item as the current recommendation and the second item as representative of the user's desired goal. The prompt contains the item descriptions and few shot examples. We select at random one of the generated critiques to inpaint into the dataset for that pair. For pairs that do not meet the similarity criteria, REGEN contains a sentinel placeholder. Dataset users can decide to train on the placeholder, to model the case where end-users do not provide critiques emulating a new search query, or to replace them with other critiques. For the “Clothing” dataset, REGEN includes LLM-generated critiques for about 18.6% of adjacent items appearing in user sequences. For “Office” dataset there are about 17.6%.

3.2.4 Narrative Generation: We aim to generate diverse narratives for conversational recommender systems, varying in:

- **Contextualization:** Narratives based on both recent and aggregate contextual information (e.g., user summaries vs explanations, endorsements of the most recent purchase).
- **Length:** Short-form and long-form narratives for varied conversational scenarios.

Using LLMs as automated evaluators, we assess generated critiques and narratives across multiple attributes. We demonstrate our approach using the Amazon product review dataset [34] and evaluate with Gemini Pro. Detailed descriptions and examples of the narratives are summarized in Table 2.

3.3 Evaluation using an Auto Rater

We employed Gemini Pro LLM as an auto-rater. The rater received the task, conversation history, generated outputs, and was prompted to score outputs based on defined attributes. In this section, we show evaluation metrics and method for Office Products dataset. We have seen similar behaviors for the larger Clothing dataset overall, which are excluded due to space constraints.

3.3.1 Overall Approach. To enhance reliability, we utilized an ensemble rating approach. Each output was rated seven times, and the scores were aggregated into a final score. We use a majority rule, where if over half of the ensemble runs agree on a score, that score becomes the final score. In the absence of a majority the average score (rounded down) is used. While computationally demanding, this method yielded more robust attribute scores. Score distributions across all conversations are presented, and corresponding user scores are included in the released dataset. Furthermore, given that new attribute fields are in-painted across the entire user sequence, we captured score distributions by randomly selecting attribute values within each user's sequence and aggregating across all users. This approach provides a comprehensive view of the attribute score distribution over the generated sequences. The definitions of attributes are summarized in Table 3.

3.3.2 Evaluation Results. For narratives, we prompt the rater LLM to score on three attributes (veracity, personalization and persuasiveness) using a seven point likert scale. All narratives are evaluated for veracity and personalization, while product endorsement is also evaluated for persuasiveness. Further, we conduct the evaluation with three calls to the LLM, one per narrative group i) Purchase reason and explanation ii) User Summaries and iii) Product endorsement. This approach is used to avoid any correlation caused by Chain of Thought (CoT) bias across different narrative types.

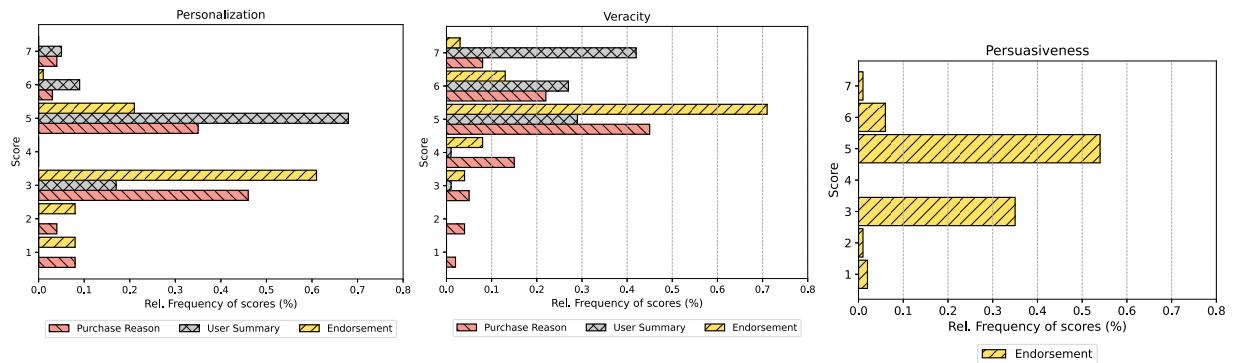
Figure 2 shows the rating distribution grouped by attributes. Summaries show higher personalization scores, while the other contextual narratives show lower scores. This is expected as personalizing to a local item context is inherently more difficult. For veracity, which checks for hallucinations and grounding, scores are relatively higher across all narratives. Persuasiveness is evaluated for product endorsement and shows some distribution at lower scores (~35% at score=3), which correlates to lower values of personalization for endorsement in some users, that can be attributed to a new product in the sequence. Overall, these results show that

Table 2: Generated narratives and their descriptions, with examples.

Context	Generated output	Description	Example
Recent/Item	Product Endorsements	A tailored endorsement/sales pitch crafted based on purchase history and reviews	"This Michael Kors Bedford Top Zip Pocket Tote Bag is a luxurious and functional choice, perfect for someone who appreciates quality and style. Based on your previous purchases of comfortable clothing and stylish sandals, this bag complements your taste for both everyday practicality and elegant accents. The durable leather, spacious interior with multiple pockets, and adjustable handles make it ideal for carrying your essentials in a sophisticated way. It's a great investment in a timeless piece that will elevate your everyday looks and be a reliable companion for years to come."
	Purchase Reasons	Concise explanations of the reasons behind a product recommendation.	"The customer wanted a stylish, relatively small tote bag with multiple pockets for organization."
	Purchase Reason Explanations	Detailed and elaborate explanations of a purchase decision of the most recent item in context of the entire history.	"The customer had previously purchased a dress and was looking for a stylish handbag that would match. They mentioned the bag was smaller than expected, but was still stylish and had multiple pockets. This suggests the customer is looking for a functional and fashionable handbag."
Aggregate/User	Brief User Summaries	Concise summaries of a user's preferences and purchase history.	"This customer is a woman who enjoys shopping for both stylish and comfortable clothing. She is also interested in fashion accessories such as handbags."
	Detailed User Summaries	Comprehensive summaries of a user's preferences and history.	"This customer is a woman who has a taste for stylish and comfortable clothing. She has purchased sandals, dresses, and underwear, all items that suggest a preference for fashion and comfort. Her reviews also mention that she values quality, especially in clothing items like dresses. She is also interested in fashion accessories such as handbags. She seems to be an occasional buyer with a more practical and less frequent shopping approach."
	User Profiles	A short phrase describing the type of user in product/shopping context.	"Fashionable and practical woman"

Table 3: Rating attributes generated by the auto LLM rater to evaluate quality of critiques and narratives.

Category	Rating Attribute	Description
Critiques	Quality	Effectiveness of the refinement query in guiding the recommender towards desired Destination Item effectively.
	Specificity	Precision of the query in narrowing the search.
	Relevance	Meaningfulness and typicality of the query's attributes for finding the desired item.
Narratives	Veracity	Accuracy and truthfulness of the purchase reason, analyzing specific evidence and inconsistencies within the user's purchase history.
	Personalization	Extent to which the narrative prioritizes information directly from the user's reviews and statements, as opposed to relying on product descriptions or general assumptions.
	Persuasiveness	LLM's ability to act as a skilled salesperson, crafting language that effectively promotes a product to the user.

**Figure 2: Narratives evaluated by rated attributes i) Personalization ii) Veracity and iii) Persuasiveness for Office Products**

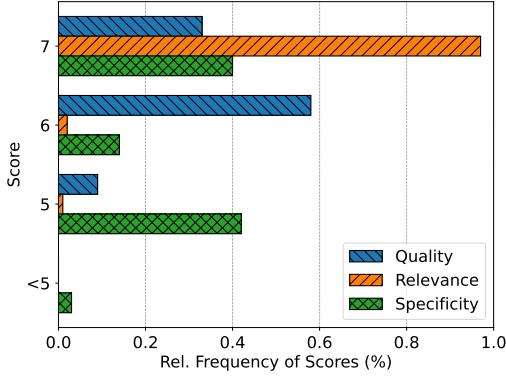


Figure 3: Critiques evaluated for i) Relevance ii) Specificity and iii) Quality for Office Products

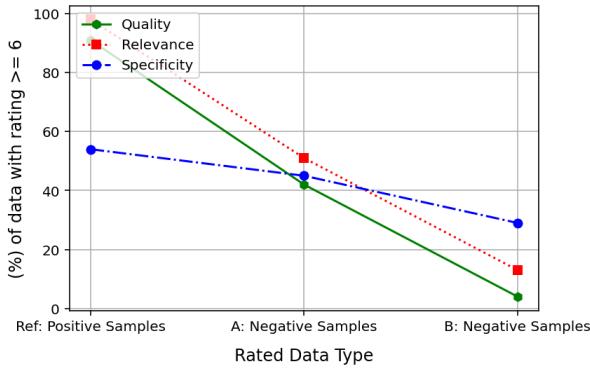


Figure 4: Steering Evaluation Bias with Negative Sampling on Office Products. Given a reference tuple (History, Critique, Next Item) A) Randomizes the History and B) randomizes the (History, Next Item) to a different user. This shows that scores are much lower on negatives and rules out positive bias in the auto rater.

in this dataset, recommendations and narratives for certain users are influenced by popular behavior that their own history.

Critiques, formulated as short queries, were evaluated on three key attribute: i) Quality, ii) Specificity and iii) Relevance, measuring their effectiveness in guiding the recommender towards the next item, given the preceding conversation. Figure 3 shows the distribution of scores, with higher than average scores across all the attributes.

3.3.3 Evaluation Bias. Figure 3 shows high auto-rater scores across all attributes for critiques. To address potential positive bias, we generated negative samples: A) by swapping the history associated with a recommended item (resetting to a search scenario) and B) by swapping both the history and item (selecting them from a different user). Actual history, critique, item and corresponding negative samples are represented as (\tilde{H}, C, I) and $(\tilde{H}, \tilde{C}, \tilde{I})$ respectively. Figure 4 shows the ratio of high scores (≥ 6), which are lower for fully randomized samples (B) and slightly higher for history-swapped

samples (A), as expected. Specificity demonstrates lower correlation to history and item, as it measures the precision of the critique itself to some extent.

4 Benchmarking Methodology

This section details the task definitions and modeling architecture.

4.1 Task Definitions

We formalize the conversational recommendation task within the framework of user interaction sequences. Consider a dataset of sequences $S = \{i_1, i_2, \dots, i_n\}$, where each interaction ($i \in \mathcal{I}$) is represented by the tuple $(ID, T, P, C, \mathcal{T})$. Here, ID denotes the unique item identifier, $T = \{t^1, t^2, \dots, t^m\}$ represents a set of natural language components describing the item (including title, detailed product description, category, and other textual metadata), P represents user-specific personalized text associated with the item (e.g., reviews, ratings, or other user-generated content), C represents an explicit user's critiquing command influencing *subsequent* item selection and \mathcal{T} is the narrative corresponding to the interactions up to n , including the current item. We define the following tasks:

- (1) **Recommendation Task (Next item prediction):** Predict the next item i_{n+1} given the user's interaction history S . This is the standard sequential recommendation objective.
- (2) **Recommendation Task with Critiquing:** Predict the next item i_{n+1} given the user's history S and user critique C .
- (3) **Conversational Recommendation Task:** Generate a recommendation i_{n+1} and natural language narrative \mathcal{T}_{n+1} , conditioned on the preceding interactions S and critiques C . This task emphasizes capturing user preferences, critiques and generating recommendations along with contextually relevant narratives.

This paper centers on the second and third tasks, as they align directly with the proposed dataset. The first task has been extensively studied, and we include it as a baseline.

4.2 Model Architectures

We benchmark two architectures, illustrated in Figure 5: a hybrid sequential recommender augmented with an LLM for generative outputs, and LUMEN, a purely LLM-based generative recommender. These are described in detail below.

4.2.1 Hybrid Sequential Recommender + LLM: This approach integrates a traditional sequential recommender with an LLM, depicted in Figure 5a. The sequential recommender focuses on modeling the collaborative filtering signal (encoded in the user-item interaction history), while the LLM is responsible for deeper semantic understanding and narrative generation. The sequential recommender produces a recommendation given user history, including content metadata for each item. We use FLARE [14] as the hybrid sequential recommender, which represents each item using both a learned embedding and a content-based encoding of the metadata (e.g. item title, description) using a language model. A perceiver resampler [17] combines the two to yield a single embedding for each item. This architecture allows steering with natural language within a sequential recommendation task. The sequence is passed through a standard transformer stack, trained using standard masked-language-model (e.g. BERT4Rec [43]). The output

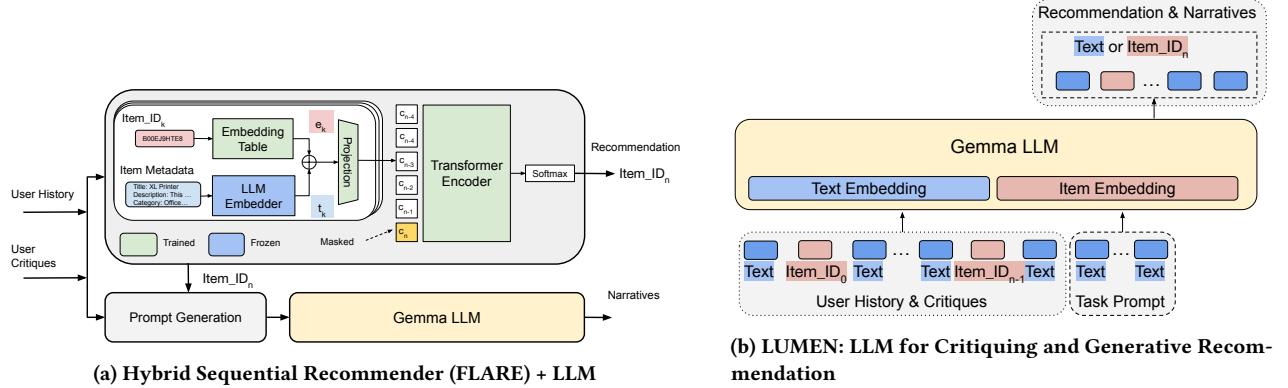


Figure 5: Modelling Architectures used for Benchmarks

recommendation is both shown to the user and conditions the LLM component. The LLM generates narratives, conditioned also on the user’s history. This architecture emulates systems that integrate conventional recommenders with LLMs, which can be trained together or separately. In our experiments, we train these components separately. To generate narratives, we prompt the LLM with the user history, consisting of a sequence of interactions, and any user critiques. The following structure is used to represent each item in the user’s purchase history:

Input Prompt: This is the summary of a user’s purchase history. First item purchase is START ID: <id> Title: <title> Category: <category> Description: <description> Price: <price> Rating: <rating> User Review: <review> User Critique: <user critique> END... Next item purchase is START ... END ... The last item purchase is: <id> (from recommender) Predict the <narrative type> for the last purchased item.

Target: <narrative>

4.2.2 LUMEN: Critiquing, Generative Recommendation and Narratives with LLM: Figure 5b illustrates LUMEN, a modeling framework where the LLM is trained to both comprehend critiques and function as a generative recommender, in addition to its narrative generation capabilities. This is achieved by augmenting an additional embedding table in the language model to enable direct generation of item ID tokens. Specifically, the new embedding table, learned through softmax loss on item vocabulary, is utilized to predict these ID tokens. Our approach draws inspiration from SemanticIDs [39], which represent items as token sequences derived from semantic embeddings and refined during LLM training. However, while SemanticIDs employs precomputed hierarchies in tokenization, LUMEN utilizes randomly initialized embeddings and single-token IDs, which are fully trained alongside the LLM. We find that this simplification works well given the vocabulary sizes of our benchmark datasets, and allows training a single end-to-end model which improves reproducibility of the benchmarks.

To manage the generation of both item IDs and language tokens, we maintain separate vocabularies. During inference, the decoder dynamically selects the appropriate vocabulary, using a specific token to signal item ID generation. The prompt structure is consistent

with the previous approach, with the exception that a placeholder token (ID_TOKEN) is used to denote the position of the item ID. This approach is analogous to the techniques employed for multi-modal inputs and outputs in Gemma [45]. The input and output prompts used are structured as follows:

Input Prompt: This is the summary of a user’s purchase history. First item purchase is START ID: ID_TOKEN Title: <title> Category: <category> Description: <description> Price: <price> Rating: <rating> User Review: <review> User Critique: <user critique> END. Next item purchase is START ID_TOKEN... END... The next item purchase, and the product endorsement are:

Target: ID: ID_TOKEN Text: <narrative>

During training, the ID_TOKEN placeholders are replaced with their corresponding item ID embeddings from the item vocabulary. During decoding, we switch from text to item vocab to generate a ID_TOKEN when a START token is detected. Once the ID_TOKEN is generated, we switch back to the text vocabulary and then continue autoregressive decoding of text tokens to generate the associated narrative. Our hypothesis is that this method enables the LLM to learn a meaningful item embedding space within the context of conversational recommendation, which is aligned with the semantic space represented by the text. We evaluate this hypothesis by comparing the performance of these two architectures across a series of benchmarks.

5 Experiments

For benchmarking, we implemented two model architectures: i) the FLARE sequential recommender [14] for recommendation generation, coupled with the Gemma 2B [45] LLM for narrative generation conditioned on the recommended item; and ii) LUMEN, utilizing Gemma 2B as the LLM, for joint recommendation and narrative generation. We evaluated these models on the REGEN dataset, focusing on the conversational tasks detailed in Section 4.1. Our evaluation spanned two categories: i) *Office Products*, comprising ~28k items, and ii) *Clothing, Shoes, & Jewelry*, comprising ~376k items. This allowed us to assess model performance across datasets with significantly different item vocabulary sizes.

Table 4: Office Products Dataset - Benchmarks with Joint Recommendation and Generation Tasks. To evaluate upper bound performance, we provide the true next item ID_{t+1} when generating Flare+LLM narratives.

Conversational Task	Metrics	Purchase Reason		User Summary		Product Endorsement		Purchase Reason Expl.		Long User Summary	
		Flare+LLM	LUMEN	Flare+LLM	LUMEN	Flare+LLM	LUMEN	Flare+LLM	LUMEN	Flare+LLM	LUMEN
Hist. ↓	Recall@10	0.124	0.098	0.124	0.10	0.124	0.10	0.124	0.091	0.124	0.098
	NDCG@10	0.0976	0.078	0.0976	0.08	0.0976	0.08	0.0976	0.071	0.0976	0.078
	MRR	0.0925	0.072	0.0925	0.074	0.0925	0.074	0.0925	0.064	0.0925	0.072
ID + Narrative	BLEU	19.99	5.43	15.05	12.47	16.77	5.34	20.2	10.34	15.14	11.4
	ROUGE	41.3	17.3	34.9	31.8	31.2	19.9	40.2	27.5	31.7	26.4
	Sem. Sim.	0.833	0.642	0.87	0.849	0.927	0.742	0.88	0.743	0.905	0.858
Hist. + Critique ↓	Recall@10	0.1402	0.123	0.1402	0.117	0.1402	0.124	0.1402	0.122	0.1402	0.122
	NDCG@10	0.113	0.102	0.113	0.098	0.113	0.103	0.113	0.1	0.113	0.1
	MRR	0.1076	0.095	0.1076	0.092	0.1076	0.097	0.1076	0.094	0.1076	0.094
ID + Narrative	BLEU	19.0	6.21	15.46	12.52	16.19	5.61	20.15	11.09	15.89	11.98
	ROUGE	41.3	19.0	36.5	31.9	30.7	20.0	40.0	28.8	33.3	27.8
	Sem. Sim.	0.816	0.651	0.87	0.851	0.929	0.747	0.891	0.757	0.897	0.87

Table 5: Clothing Dataset - Benchmarks with Joint Recommendation and Generation Tasks. To evaluate upper bound performance, we provide the true next item ID_{t+1} when generating Flare+LLM narratives.

Task	Metrics	Purchase Reason		User Summary	
		Flare+LLM	LUMEN	Flare+LLM	LUMEN
Hist. ↓	Recall@10	0.126	0.127	0.126	0.118
	NDCG@10	0.117	0.111	0.117	0.100
	MRR	0.115	0.106	0.115	0.094
ID + Narrative	BLEU	19.2	10.2	16.4	12.8
	ROUGE	39.8	29.7	33.9	29.9
	Sem. Sim.	0.89	0.78	0.92	0.90
Hist. + Critique ↓	Recall@10	0.136	0.143	0.136	0.151
	NDCG@10	0.124	0.122	0.124	0.134
	MRR	0.122	0.115	0.122	0.128
ID + Narrative	BLEU	19.2	10.5	16.3	13.3
	ROUGE	40.0	30.2	34.1	30.3
	Sem. Sim.	0.89	0.79	0.92	0.90

5.1 Setup

5.1.1 Training and Evaluation. We assessed recommendation performance using standard metrics: Recall@N, NDCG@N and MRR, with N set to 10. These metrics are widely adopted in recommendation tasks. For evaluating the quality of generated text, we employed BLEU and ROUGE to measure n-gram similarity. Additionally, we utilized a Semantic Similarity Metric, calculated as the cosine similarity between sentence embeddings generated by the Gecko embedding model [22]. For partitioning the dataset into training, validation, and test sets, we adopted the leave-one-out approach [20], a common practice in sequential recommendation. Specifically, the first $N - 2$ interactions were used for training, the first $N - 1$ interactions with the last item held out were used for validation, and all N interactions with the last item held out were used for testing.

5.1.2 Implementation Details. Our primary objective in benchmarking is to provide representative performance results using well-known architectures from recent research for conversational

recommendation, rather than to advocate for a specific model configuration. While we added some variations that work well for the task and the dataset, and performed basic hyperparameter tuning within each framework, we acknowledge that more extensive optimization, including finer-grained tuning and the use of larger component models, could potentially yield improved results. We leave these for future investigation.

With the hybrid sequential recommender, we employed a Masked Language Modeling (MLM) task during training to enhance sample efficiency. We used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and adopted the hyperparameter configurations detailed in [14]. Further, we made a simplifying assumption that the ground truth item is available to the LLM. Modeling the error in item recommendation would require co-training or evaluating the sequential encoder and decoder LLM models together, which is difficult to reproduce. On the other hand, with this assumption, we view the results with this approach as an upper bound for the joint recommendation and generation tasks. The approach is also easily reproducible, allowing for comparisons of other component models.

With LUMEN, utilizing the LLMs for recommendation, we trained on subsequences up to length $N - 2$ for the joint recommendation and generation task. We use the Gemma 2B model with a batch size of 512 and a constant learning rate of 1e-5, training for up to 10,000 steps with early stopping.

5.2 Results and Discussion

Our experimental results, summarized in Table 4 and Table 5, lead to the following key findings:

5.2.1 Recommendation Performance. To establish baselines, we first compare recommendation metrics, which are well-studied for several recommender models in related work on amazon datasets.

On REGEN Clothing (Table 5), compared to FLARE (See [14] and references within) as a baseline, LUMEN trained on joint task achieves competitive performance despite generating recommendations and narratives within a single model, often achieving within ± 0.01 Recall@10 and MRR, and improving performance in some cases.

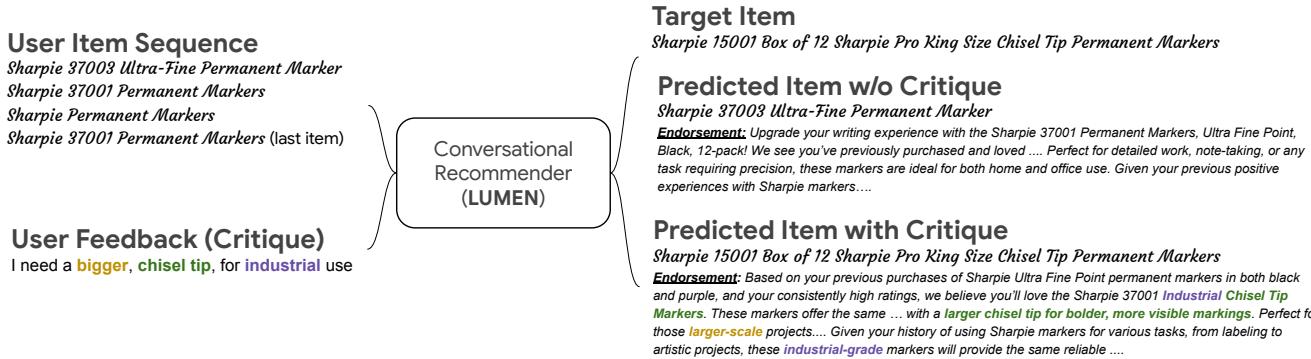


Figure 6: An example trace from LUMEN, using a joint generative task on REGN, demonstrates the conversational recommender model’s ability to process purchase interactions and in-context natural language critiques. This processing improves the predicted item and generates a natural language endorsement that aligns with the prediction.

For Office Products (Table 4) however, we see slightly lower recommendation metrics, within 10–15% of the baseline. The Office dataset has a significantly smaller vocabulary ($\approx 27k$) and training data size ($\approx 90k$) (Table 1). We believe this indicates that a sufficiently large training dataset is required to retain the recommendation performance in models trained on joint tasks.

5.2.2 Recommendation Performance with Natural Language Critiques. Incorporating user critiques consistently enhances recommendation performance, with improvements of +0.01–0.02 in Recall@10 and NDCG@10 and +0.01–0.03 in MRR observed for both hybrid and end-to-end LUMEN models (Table 5).

5.2.3 Narrative Performance. As expected, LUMEN’s narrative quality scores (BLEU, ROUGE) are lower than those of a hybrid model conditioned directly on ground-truth items. We attribute this performance gap primarily to recommendation errors, as LUMEN’s narratives are based on its own predicted items in a single end to end model. We find that this dependency was less impactful for aggregate narratives such as “User Summary,” which rely less on the accuracy of the most recent item recommendation and more on the overall user history. In this case, LUMEN compares well with an LLM trained to generate narratives alone. Further, we expect that an integrated hybrid architecture where the output of the recommender is directly used as input to LLM is likely to see similar errors.

5.2.4 Validation with Model Traces. Figure 6 shows an example trace demonstrating recommendation, natural language understanding and generation capabilities of the LUMEN model trained on the joint tasks with the REGN dataset.

6 Limitations and Conclusion

This paper introduced REGN, a novel dataset enhancing Amazon Reviews with personalized narratives and critiques, facilitating a new task in conversational recommendation. While a comprehensive human evaluation of REGN’s critiques for “naturalness” would be beneficial, especially given their intended role as user feedback proxies, we argue that REGN remains valuable even without

it. Obtaining reliable human annotations for subjective, context-dependent textual tasks at scale is challenging. Moreover, the utility of a dataset is not solely dependent on perfect human alignment, as evidenced by the widespread use of Amazon Reviews despite its limitations. We demonstrate REGN’s value by showing its effectiveness in developing conversational recommendation models that integrate sequential recommendations, critique generation, and narrative generation.

Our benchmarking efforts, while effective, utilized relatively simple architectures, including randomly initialized embeddings and single-token IDs. For larger item spaces, future work could explore optimizations like multi-token semantic IDs [39] or other richer representations.

In summary, REGN provides a dataset with consistent user preferences, recommendations, and generated narratives, enabling the study of LLM capabilities in conversational recommendation. We evaluated REGN using LUMEN, an LLM-based model for joint recommendation and narrative generation, demonstrating its utility, along with sequential recommender models. We believe REGN serves as a fundamental resource for studying the capabilities of conversational recommender models, a crucial step towards personalized multi-turn systems.

References

- [1] Krisztian Balog, Filip Radlinski, and Alexandros Karatzoglou. 2021. On Interpretation and Measurement of Soft Attributes for Recommendation. In *SIGIR ’21*. ACM, New York, NY, USA, 890–899.
- [2] Robin Burke, Kristian Hammond, and Benjamin Christopher Young. 1996. Knowledge-Based Navigation of Complex Information Spaces. In *AAAI 96 and IAAI 96*, Vol. 1. 462–468.
- [3] Chanapa Chanarong, Chawisa Paosirikul, Saranya Maneeroj, and Atsuhiko Takasu. 2022. HybridBERT4Rec: A Hybrid (Content-Based Filtering and Collaborative Filtering) Recommender System Based on BERT. *IEEE Access* 10 (2022), 56193–56206.
- [4] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Model. User-adapt Interact.* 22, 1 (April 2012), 125–150.
- [5] Tao Chen, Siqi Zuo, Cheng Li, Mingyang Zhang, Qiaozhi Mei, and Michael Bendersky. 2024. Unlocking the ‘Why’ of Buying: Introducing a New Dataset and Benchmark for Purchase Reason and Post-Purchase Experience. *arXiv:2402.13417 [cs.IR]* <https://arxiv.org/abs/2402.13417>
- [6] Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog Impainting: Turning Documents into Dialogs. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan

- Sabato (Eds.). PMLR, 4558–4586. <https://proceedings.mlr.press/v162/dai22a.html>
- [7] Sumanth Doddapaneni, Krishna Sayana, Ambarish Jash, Sukhdeep Sodhi, and Dima Kuzmin. 2024. User Embedding Model for Personalized Language Prompting. arXiv:2401.04858 [cs.CL] <https://arxiv.org/abs/2401.04858>
- [8] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open* 2 (Jan. 2021), 100–126.
- [9] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). In *RecSys '22*. Association for Computing Machinery, New York, NY, USA, 299–315.
- [10] Sally Goldman and Yuri Vasilevski. 2024. RoboRater: Automating Ratings of Task-Oriented Conversations using LLMs. In *AAAI 2024 Spring Symposium on User-Aligned Assessment of Adaptive AI Systems*. https://aaир-lab.github.io/aia2024/papers/goldman_aia24.pdf
- [11] Peter Grasch, Alexander Felfernig, and Florian Reinfrank. 2013. ReComment: towards critiquing-based recommendation with speech interaction. In *RecSys '13*. Association for Computing Machinery, New York, NY, USA, 157–164.
- [12] Shuyu Guo, Shuo Zhang, Weiwei Sun, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2023. Towards Explainable Conversational Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) (*SIGIR '23*). Association for Computing Machinery, New York, NY, USA, 2786–2795. doi:10.1145/3539618.3591884
- [13] Jesse Harte, Wouter Zorgdrager, Panos Louridas, Asterios Katsifodimos, Dietmar Jannach, and Marios Fragkoulis. 2023. Leveraging Large Language Models for Sequential Recommendation. In *RecSys '23*. Association for Computing Machinery, New York, NY, USA, 1096–1102.
- [14] Liam Hebert, Marialena Kyriakidi, Hubert Pham, Krishna Sayana, James Pine, Sukhdeep Sodhi, and Ambarish Jash. 2024. FLARE: Fusing Language Models and Collaborative Architectures for Recommender Enhancement. arXiv:2409.11699 [cs.IR] <https://arxiv.org/abs/2409.11699>
- [15] Yupeng Hou, Jiacheng Li, Zhanhui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging Language and Items for Retrieval and Recommendation. arXiv:2403.03952 [cs.IR] <https://arxiv.org/abs/2403.03952>
- [16] Wenyue Hua, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2023. How to Index Item IDs for Recommendation Foundation Models. In *SIGIR-AP '23*. Association for Computing Machinery.
- [17] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General Perception with Iterative Attention. In *ICML '21 (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 4651–4664.
- [18] Dietmar Jannach, Althash Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* 54, 5 (May 2021), 1–36.
- [19] Meihuizi Jia, Ruixue Liu, Peiying Wang, Yang Song, Zexi Xi, Haobin Li, Xin Shen, Meng Chen, Jinhui Pang, and Xiaodong He. 2022. E-ConvRec: a large-scale conversational recommendation dataset for E-commerce customer service. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 5787–5796.
- [20] Wang-Cheng Kang and Julian McAuley. 2018. Self-Attentive Sequential Recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. 197–206. doi:10.1109/ICDM.2018.00035
- [21] Sein Kim, Hongseok Kang, Seungyoon Choi, Donghyun Kim, Minchul Yang, and Chanyoung Park. 2024. Large Language Models meet Collaborative Filtering: An Efficient All-round LLM-based Recommender System. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona, Spain) (*KDD '24*). Association for Computing Machinery, New York, NY, USA, 1395–1406. doi:10.1145/3637528.3671931
- [22] Jinhyuk Lee, Zhuyun Dai, Xiaorui Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. 2024. Gecko: Versatile Text Embeddings Distilled from Large Language Models. arXiv:2403.20327 [cs.CL] <https://arxiv.org/abs/2403.20327>
- [23] Megan Leszczynski, Shu Zhang, Ravi Ganti, Krisztian Balog, Filip Radlinski, Fernando Pereira, and Arun Tejasvi Chaganty. 2023. Talk the Walk: Synthetic Data Generation for Conversational Music Recommendation. arXiv:2301.11489 [cs.IR] <https://arxiv.org/abs/2301.11489>
- [24] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*.
- [25] Xiangnan Li, Lei Zhao, Xiang He, Mengdi Zhang, Qi Wu, and Panupong Pasupat Chang. 2017. Redial: A large-scale dataset for group recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 1157–1160.
- [26] Yaoyiran Li, Xiang Zhai, Moustafa Alzantot, Keyi Yu, Ivan Vulic, Anna Korhonen, and Mohamed Hammad. 2024. CALRec: Contrastive alignment of generative LLMs for sequential recommendation. In *RecSys 2024*.
- [27] Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. 2024. LLM-REDIAL: A Large-Scale Dataset for Conversational Recommender Systems Created from User Behaviors with LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 8926–8939. doi:10.18653/v1/2024.findings-acl.529
- [28] Cheng-Te Lin, Hung-Yi Lee, Ming-Feng Sun, Wen-Chi Hwang, et al. 2021. M6-Rec: Generative Pretrained Language Models for Conversational Recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 629–637.
- [29] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 4335–4347. doi:10.18653/v1/2021.emnlp-main.356
- [30] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards Conversational Recommendation over Multi-Type Dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1036–1049. doi:10.18653/v1/2020.acl-main.98
- [31] Kai Luo, Scott Sanner, Ga Wu, Hanze Li, and Hojin Yang. 2020. Latent Linear Critiquing for Conversational Recommender Systems. In *WWW '20*. Association for Computing Machinery, New York, NY, USA, 2535–2541.
- [32] Kai Luo, Hojin Yang, Ga Wu, and Scott Sanner. 2020. Deep Critiquing for VAE-based Recommender Systems. In *SIGIR '20*. Association for Computing Machinery, New York, NY, USA, 1269–1278.
- [33] Preksa Nema, Alexandros Karatzoglou, and Filip Radlinski. 2021. Disentangling Preference Representations for Recommendation Critiquing with β -VAE. In *CIKM '21*. Association for Computing Machinery, New York, NY, USA, 1356–1365.
- [34] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 188–197.
- [35] Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. 2024. User-LLM: Efficient LLM Contextualization with User Embeddings. arXiv:2402.13598 [cs.CL] <https://arxiv.org/abs/2402.13598>
- [36] Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. 2024. User-LLM: Efficient LLM Contextualization with User Embeddings. arXiv:2402.13598 [cs.CL]
- [37] Zaopeng Qiu, Xian Wu, Jingyu Gao, and Wei Fan. 2021. U-BERT: Pre-training user representations for improved recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4320–4327.
- [38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.
- [39] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, Maciej Kula, Ed Chi, and Maheswaran Sathiamoorthy. 2023. Recommender Systems with Generative Retrieval. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 10299–10315. https://proceedings.neurips.cc/paper_files/paper/2023/file/20dcab0f14046a5c6b02b61da9f13229-Paper-Conference.pdf
- [40] Ahmed Rashed, Sherene Elsayed, and Lars Schmidt-Thieme. 2022. Context and Attribute-Aware Sequential Recommendation via Cross-Attention. In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) (*RecSys '22*). Association for Computing Machinery, New York, NY, USA, 71–80. doi:10.1145/3523227.3546777
- [41] Hideo Shimazu. 2002. ExpertClerk: A Conversational Case-Based Reasoning Tool for Developing Salesclerk Agents in E-Commerce Webshops. *Artificial Intelligence Review* 18, 3 (Dec. 2002), 223–244.
- [42] Hakim Sidahmed, Sameer Ahuja, and Manoj Kumar Tiwari. 2022. Generating personalized recommendations via large language models (llms). (2022).
- [43] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *CIKM '19*. Association for Computing Machinery, New York, NY, USA, 1441–1450.
- [44] Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530 [cs.CL] <https://arxiv.org/abs/2403.05530>
- [45] Gemma Team. 2024. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295 [cs.CL] <https://arxiv.org/abs/2403.08295>
- [46] Cynthia A Thompson, Mehmet H Göker, and Pat Langley. 2004. A personalized system for conversational recommendations. *J. Artif. Intell. Res.* 21, 1 (March 2004), 393–428.
- [47] Bowen Wang, Zhongkai Lin, Rui Zhang, Hong Yang, Yan Song, Dongyan Zhang, Tao Yu, Bing Zhou, and Qiang Yang. 2024. Bridging Language and Items

- for Retrieval and Recommendation. *arXiv preprint arXiv:2403.03952* (2024). arXiv:2403.03952 [cs.IR]
- [48] Ga Wu, Kai Luo, Scott Sanner, and Harold Soh. 2019. Deep language-based critiquing for recommender systems. In *RecSys '19*. Association for Computing Machinery, New York, NY, USA, 137–145.
- [49] Yuhui Zhang, Hao Ding, Zeren Shui, Yifei Ma, James Zou, Anoop Deoras, and Hao Wang. 2021. Language models as recommender systems: Evaluations and limitations. (2021)
- [50] Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Adapting Large Language Models by Integrating Collaborative Semantics for Recommendation. In *ICDE '24*. IEEE.
- [51] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information knowledge management*. 1893–1902.
- [52] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards Topic-Guided Conversational Recommender System. arXiv:2010.04125 [cs.CL] <https://arxiv.org/abs/2010.04125>
- [53] Yaochen Zhu, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. 2024. Collaborative large language model for recommender systems. In *The ACM Web Conference (WWW) '24*. Association for Computing Machinery, New York, NY, USA.
- [54] Jianhuan Zhuo, Jianxun Lian, Lanling Xu, Ming Gong, Linjun Shou, Daxin Jiang, Xing Xie, and Yinliang Yue. 2022. Tiger: Transferable interest graph embedding for domain-level zero-shot recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2806–2816.